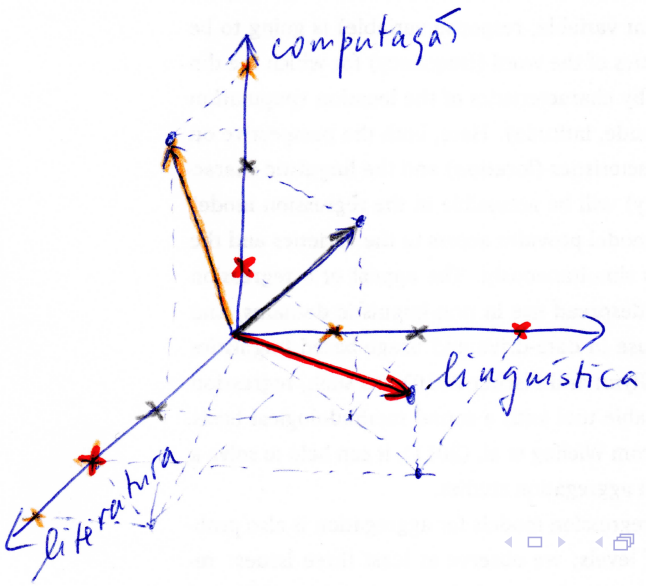
What can you do with 100 novels?

Reflections on a European distant reading initiative and beyond

Diana Santos

d.s.m.santos@ilos.uio.no



COST "Distant reading for European literary history"

Distant Reading

This action will affect the way scholars in the Humanities do research, but also the way institutions like libraries will make their holdings available to researchers in the future

- http://www.cost.eu/COST_Actions/ca/CA16204
- https://www.distant-reading.net/
- https://distantreading.github.io/

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

A little more about this COST action

Columbano Bordalo Pinheiro (1857-1929): O Grupo do Leão, 1885

- Nov. 2017 - Oct. 2021
- Three working groups:
  - WG1 building the ELTeC collection;
  - WG2 computational tools;
  - WG3 literary analysis

- The several collections:
  `https://distantreading.github.io/ELTeC/`

ELTeC collection: some problems

Period 1840-1919

- different cultures, different literary histories
- how to select (or find) 100 novels?
- different realities – how much should one use a one-size-fits-all design (novel length, women authors, different periods, canonicity)?
- how to harmonize linguistic and NLP tools for different languages

Ernst Ludwig Kirchner

Theodor Kittelsen

José Malhoa

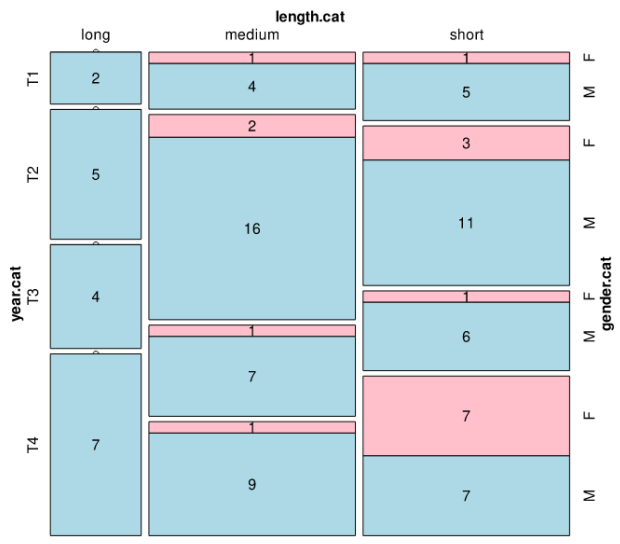
The Portuguese collection (now)

- 100 works, released on 13 November 2020
- Opportunistic selection - whatever was available, until the 100 (in fact 118)
- One third were historical novels!
- Very difficult to obtain long novels (as opposed to German, for example)
- Only 14 woman writers
- Most long novels were canonical

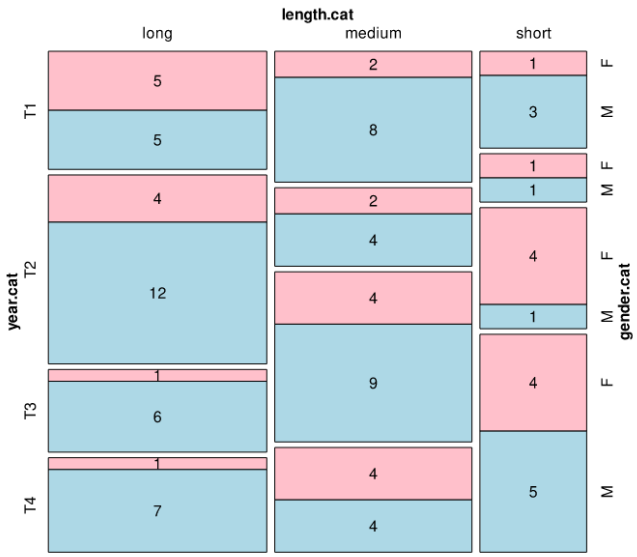
Digitization projects in Portugal (and Brazil) seemed to concentrate on canonical works.

Parallel does not mean exactly the same...

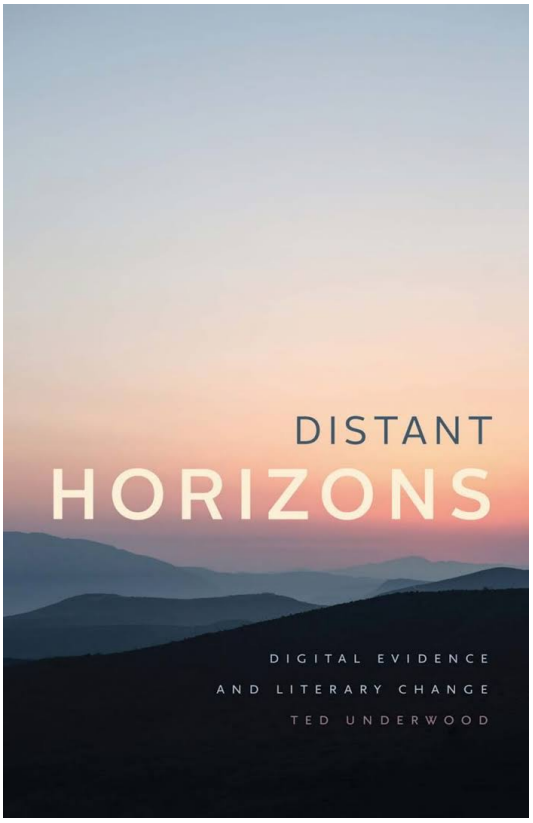
Portuguese



German



Distant reading



- large-scale research
- not necessarily to save the forgotten / archive (Moretti's moralism)
- another way to look: lenses that see larger periods

  Can distant readers write quantitative literary history that is nevertheless detailed enough, streamlined enough, and lively enough to interest a wide range of readers? If we can't, then no argument will save us: what we are doing may be important, but it will belong in the social sciences.

(Underwood, 2019, p. xxii)

Size matters

The studies reported in the literature usually concern thousand, even hundreds of thousand books. So, what to do with 100 (times 15) European books?

- exploratory studies
- cross-cultural studies
- seeds for "real" distant reading: OCR improving, estimation of other quantities, etc.

Christmas in ELTeC-por

Machado de Castro (1731-1822): Presépio, Basílica da Estrela (1781-1786), detail

- Only 17 works mention at all *Natal* (Christmas), and only three novels describe events in Christmas (eve/day)
- *Páscoa* (Easter) is only mentioned in 15 works, and no scene happens then.
- *Carnaval* is mentioned in 10 novels, two of which have a scene in this period.

Fun fact: there are several regions called *Natal* (a city in Brazil, a region in South Africa, now KwaZulu-Natal) because they were discovered by Portuguese on Christmas day.

Titles of books in ELTeC-por

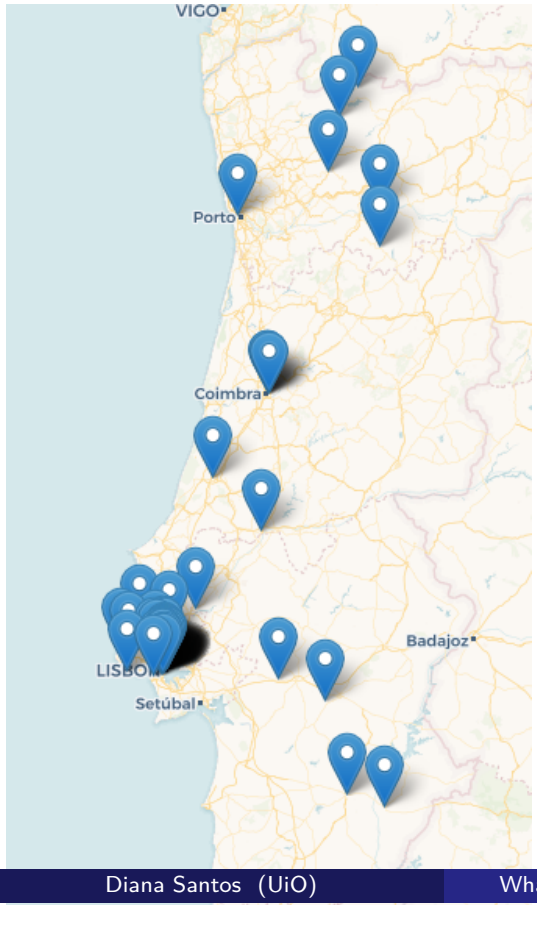
Name of a woman	7
Name of a man	14
Description of a woman	9
Description of a man	11
(Additional) mention to a woman	2
(Additional) mention to a man	8
Total Women	18
Total Men	34

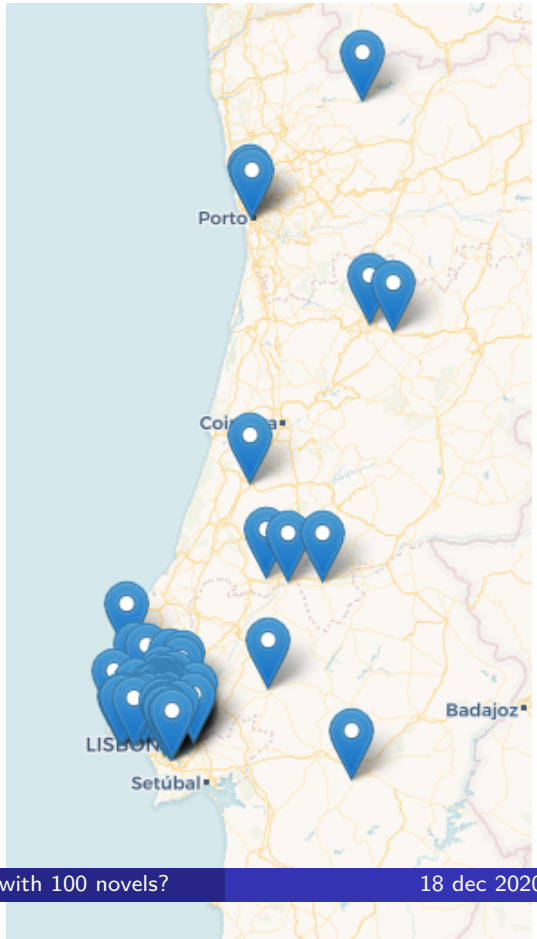
Named entities, demonyms and professions and titles

- Across colections
- Inside each collection

- Geographic distribution
- How many politicians (Napoleon) and writers (Camões)? And fictional characters?
- Historical events and nationalities/groups
- Emotions and their relations to gender, to place, etc.

Maps per authors: Eça de Queirós vs. Lobo Antunes





Preparing the future

One can also see this work as **just preparatory**, to decide how to do things in large scale

- index works by ortography (five or six different)
- develop OCR correctors (or new OCR software)
- develop normalizers (in order to use modern NLP)
- also to quantify problems in Google books

And also as an **addition to own repositories** (for Portuguese, we have currently in Linguateca ca 850 works and growing, fully parsed and undergoing diverse semantic annotation).

Obrigada!